Math 455 Class 40 May 7

Solutions to Midterm 2 are now posted.

Warning: written in a big hurry. Email me about any mistakes.

Final Exam: Thursday May 15 5:40PM
CW 331.

(This is directly from Banner – I will check again right before the exam.)

HW 13 – Sample Final Exam.

Sins of Omission:

Principal Component Analysis

Instrumental Variables

How to run a Lasso or Ridge Regression

Root Mean Square Error.

Partial Least Squares.

Reference for Theory of Regression.

Many results in Faraway are stated without proof: you can find proofs in many texts, e.g. Seber and Lee.

Alternative Error Measures: one standard measure is Root Mean Square Error.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Somewhat analagous to distance $d(y, \hat{y})$

$$d(y, \hat{y}) = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Instrumental Variables: very important if you're going on in Econ / Social Science.

Idea: uncorrelated-to-error proxy.

Example: You want to study the effect of smoking (X) on cancer (Y).

Problem: Smoking is correlated with other risky health behavior that influences Y.

(Instrumental Variables) Solution:

Find Z which is highly correlated with X but uncorrelated to the "other risky health behavior". Say Z = tobacco tax rates.

The idea is that we can study the effect of Z on Y without the problems associated to studying the effect of X.

Econ texts will have whole chapters on Instrumental Variables; Faraway does not mention them.

Warning: People may be smarter than your model. If there is a way to get around tobacco taxes, they will do it: Native American Reservations / Nations, drive to PA, etc.

Idea: you can feed tests & solutions to your favorite LLM and ask it to make up more tests/questions.

Remember: Linear Regression is a "language" and a lot of what you're doing is learning vocabulary. This is particularly true for the multiple choice questions.

This is precisely the kind of thing that the LLM does well.

Lasso regression: lars

Ridge regression: lm.ridge (MASS)

Principal Component Analysis.

Idea: find the directions of largest variation in our data; make this a the new coordinate system.

This is the same thing as the "Singular Value Decomposition" of Linear Algebra.

Recommendation: If you plan to continue in Stats / CS / Econ, good to know linear algebra very well.

Problem with PCA: Finding an interpretation of the new coordinates. This is often possible for the first few components.

Example: interest rate curves; measure interest rates at 1yr, 2yr, ..., 30 yr. 30 variables.

They are all highly correlated, so 1^{st} principal component turns out to be (in most data sets) the average. or "level"

2^{nd} and 3^{rd} components: slope & curvature.